Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A 5-miRNA blood signature for gastric cancer diagnosis and therapeutic targeting derived from tumor pathways and validated by machine learning and single-cell RNA-seq

Sorour Hassani
Semnan University
Maryam Momeni
University of Isfahani
Maryam Siahtiri
University of Kashan
Sina Massahi
University of Tabriz
Shayan Jalali
University of Eastern Piedmont/Piemonte Orientale
Saeedeh Salehi
Tehran University of Medical Sciences
Alieh Gholaminejad
al_gholami@yahoo.com

Isfahan University of Medical Sciences

Research Article

Keywords: Gastric cancer diagnosis, miRNA biomarkers, machine learning, pathway analysis, Random Forest, Bioinformatics

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10304374/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Gastric cancer (GC) remains a leading cause of global cancer mortality, with late-stage diagnosis contributing significantly to poor patient outcomes. Circulating microRNAs (miRNAs) offer promise due to their stability in biofluids and established roles in carcinogenesis. However, existing miRNA biomarker candidates for GC suffer from inconsistent validation, limited specificity, and insufficient mechanistic links to gastric tumor biology. Establishing such mechanistic links for a miRNA signature would not only enhance its diagnostic value but also position it as a potential therapeutic target.

Methods

Differentially expressed genes (DEGs) and miRNAs (DEMs) were identified from tissue mRNA (GSE54129, GSE113255) and blood miRNA/mRNA datasets (GSE106817, GSE174302). We first identified shared dysregulated Reactome pathways between tissue DEGs and the target genes of blood-derived DEMs. miRNAs associated with these gastric cancer-specific pathways were selected as candidates. An ensemble of ten machine learning (ML) feature selection methods was then used to refine a minimal miRNA panel, which was subsequently validated using Random Forest and Naïve Bayes classifiers in independent dataset. To further validate them and explore their cellular origin, we performed single-cell RNA sequencing (scRNA-seq) on GC tissue. Additionally, a drug repurposing analysis was performed using the DGIdb database.

Results

A 5-miRNA panel (miR-124-3p, miR-23a-3p, miR-22-3p, miR-29b-3p, and miR-92a-3p) was identified. It demonstrated high diagnostic accuracy in the discovery set (AUC = 98.50%) and maintained robust performance upon external validation (AUC = 95.30%). scRNA-seq analysis confirmed the expression of target genes for a core 3-miRNA subset within key tumor microenvironment cell types (fibroblasts, endothelial, and epithelial cells). Furthermore, drug repurposing analysis identified four FDA-approved drugs (Olaparib, Doxorubicin Hydrochloride, Bithionol, and Daunorubicin Liposomal) as potential therapeutic candidates interacting with the signature's target genes.

Conclusion

This integrated multi-level transcriptomics and ML pipeline successfully discovered a high-performance, blood-based miRNA signature for GC. The signature is validated through scRNA-seq and ML, and linked to candidate therapeutics, offering a promising strategy for non-invasive diagnosis and therapeutic targeting.

1. Introduction

Gastric cancer remains a formidable global health challenge, claiming over 769,000 lives in 2020 and ranking among the top causes of cancer mortality [1]. Late diagnosis is the crux of the problem; most cases are detected at advanced stages, when treatment options dwindle, and survival rates plummet [2]. Current early detection methods, like endoscopy, are accurate but invasive and costly, limiting their use in routine screening. Blood-based biomarkers, such as CEA and CA19-9, often lack the sensitivity and specificity needed for early-stage detection [3, 4], underscoring the urgent need for better diagnostic tools.

miRNAs are small, stable, non-coding RNAs that regulate gene expression and are gaining popularity as biomarkers due to their detectability in biofluids [5]. Panels of miRNAs have already proven successful in diagnosing complex diseases and cancers [6–9], and early evidence indicates similar potential for GC using blood or gastric juice [5]. Recently, miRNAs have become recognized as components of liquid biopsy [10]; however, ensuring that these miRNAs are specific to GC tumor development remains a challenge. We addressed this by using integrated pathway analysis to link DEMs with DEGs in GC and identify molecular markers of the disease. To do this, we hypothesized that specific blood-based miRNAs for GC could be identified by focusing on disease-specific pathways. By identifying shared enriched biological pathways between circulating DEMs and tumor DEGs and extracting DEMs from these pathways, we isolated detectable blood-derived DEMs that play a key role in GC progression.

On the other hand, ML can play a pivotal role in our approach, enabling us to navigate the complexity of multi-omics data and pinpoint optimal biomolecule candidates [11–13]. Unlike traditional methods, ML can uncover subtle patterns in high-dimensional datasets, making it ideal for selecting biomarkers that are both accurate and biologically relevant. Recently, ML approaches have enabled the non-invasive, efficient, and accurate early detection of various cancers, including GC [14, 15]. Feature selection is a powerful ML technique that identifies the optimal combination of biomarkers, facilitating the discovery of highly predictive miRNA biomarker panels in blood [16]. Therefore, we used ML to refine the selection of blood-based miRNAs, yielding a biomarker panel with fewer members, making it more practical for clinical testing.

To determine which cell types in the tumor microenvironment (TME) produce and release the identified circulating miRNAs, we performed scRNAseq. Guided by our novel strategy (which identifies a blood signature by intersecting common pathways from tissue and blood transcriptomics that are fundamental to gastric cancer), this cell-type-specific analysis provides new insights into the cellular drivers of GC pathogenesis. Consequently, it not only validates the central role of the miRNA signature but also nominates promising cellular targets for future therapy. Furthermore, as our workflow identifies key driver genes within these fundamental GC pathways, the resulting biomarker signature may itself play a crucial role in pathogenesis and represent valuable drug targets. Therefore, we performed drug repurposing analysis for these signature genes. In this study, we present a novel three-step framework for discovering and validating high-specificity circulating miRNA biomarkers and drug targets GC. In Step

1, we integrated transcriptomic data from GC tissues and blood to identify DEMs. This involved identifying shared dysregulated pathways between tissue mRNA (GSE54129, GSE113255) and blood-derived miRNA target genes (GSE106817), refined via intersection with blood mRNA data (GSE174302). Moreover, finally, extracting DEMs whose targets enrich these GC-specific pathways. In Step 2, we employed a robust ML strategy, applying 10 distinct feature selection methods to DEMs from Step 1 to identify a consensus biomarker panel. The panel's performance was then validated using Naïve Bayes (NB) and Random Forest (RF) classifiers on the discovery (GSE106817) and independent validation (GSE164174) datasets. This approach bridges tissue-level pathology with circulating miRNA signals, ensuring biological relevance while leveraging computational rigor to derive a minimally invasive, high-performance diagnostic tool for gastric cancer. Finally, in Step 3, we validated the miRNA signature using scRNA-seq and performed drug repurposing.

Our workflow (Fig. 1) began with pathway analysis, continued with ML-driven miRNA selection, and culminated in scRNA-seq validation and drug repurposing, paving the way for a robust diagnostic tool and therapeutic drug.

2. Materials and Methods

2.1. Dataset Selection

This study utilized several transcriptomic datasets from the Gene Expression Omnibus (GEO) database, including two tissue datasets: GSE54129 (111 Tumor (T), 21 Normal (N)) and GSE113255 (130T, 10N); blood mRNA dataset: GSE174302 (27T, 15N); and two blood miRNA datasets: GSE106817 (115T, 2759N) and GSE164174 (1417T, 1417N). Datasets were selected based on defined inclusion criteria—only those derived from human tumor tissue and blood samples of healthy individuals and gastric cancer patients with adequate sample size, quality, and clinical relevance were included. In contrast, datasets from cell lines or non-human sources were excluded. Table 1 summarizes the sample sizes, characteristics, and specific roles of these datasets in the study.

Table 1
Transcriptomic datasets used in this study

Dataset ID	Sample type	Technology/platform/platform ID	Sample size (T/N)	Usage
GSE54129	GC tissue	Microarray/ Affymetrix/ GPL570	111/21	Pathway enrichment in step 1
GSE113255	GC tissue	RNAseq/ Illumina/ GPL18573	130/10	Pathway enrichment in step 1
GSE174302	GC plasma	RNAseq/HiSeq X Ten (Homo sapiens)/ GPL20795	27/15	Validation of miRNA targets (pathway enrichment of GSE106817 dataset) in step 1
GSE106817	GC serum miRNA	Microarray/ 3D-Gene Human miRNA V21_1.0.0/ GPL21263	115/2759	Pathway enrichment in step 1, Feature selection for finding the biomarker panel, and validation of that in step 2
GSE164174	GC serum miRNA	3D-Gene Human miRNA V21_1.0.0/ GPL21263	1417/1417	External validation of biomarker panels in step 2

2.2. Step 1: GC-Specific DEMs Discovery

2.2.1. Differential Expression Analysis

Following standard preprocessing, DEGs between cancerous and normal samples were identified from the microarray and RNA-seq datasets using the GEO2R tool. DEGs were defined based on a threshold of $|\log_2 FC| > 0.585$ ($|FC| > 1.5$) and an adjusted p-value < 0.05 to ensure statistical significance.

2.2.2. Finding GC-Specific DEMs

We identified pathways shared between circulating DEM targets and tissue DEGs, then extracted the DEMs associated with these pathways from the blood miRNA dataset.

2.2.2.1. GC Tissue Pathway Enrichment Analysis

DEGs were identified from two independent GC tissue datasets (GSE54129 and GSE113255). Overlapping DEGs between both datasets were selected to enhance reliability. Pathway enrichment analysis of these DEGs was conducted using the Reactome database via Enrichr to identify GC-specific pathways. The Benjamini-Hochberg correction was used to control the false discovery rate, and pathways with adjusted p-values < 0.05 were considered significant.

2.2.2.2. Circulating miRNA Pathway Enrichment Analysis

DEMs from the circulating miRNA dataset (GSE106817) were mapped to their target genes using miRTarBase (2025 updated version). Only strong evidence of miRNA-target interactions from miRTarBase reports was considered. To refine GC-specific miRNA target genes and ensure their specificity to GC, we identified their intersection with GC blood-derived DEGs from the GSE174302 blood dataset. Pathway enrichment analysis was then conducted on these refined targets using the Reactome database via Enrichr. The Benjamini-Hochberg correction was applied to control the false discovery rate, and pathways with adjusted p-values < 0.05 were deemed significant.

2.2.2.3. Identification of DEMs Associated with Shared Pathways Between Tumor Tissue and Blood

Shared pathways between tumor tissue DEGs and target genes of circulating DEMs were selected. Finally, DEMs whose target genes were enriched in these shared pathways in the circulating miRNA dataset (GSE106817) were selected and used for step 2.

2.3. Step 2: Biomarker Panel Selection and Validation by ML

The expression values of identified miRNAs in the previous section were extracted from the circulating miRNA dataset (GSE106817) and used as input features for ML analysis.

2.3.1. Feature Selection

To identify a clinically efficient biomarker panel, we applied 10 ML-based feature selection methods, including Weight by Gini Index, Information Gain, Information Gain Ratio, Rule, Chi-Squared Statistic, Tree Importance, Uncertainty, Deviation, Correlation, and Relief, to the GSE106817 dataset in RapidMiner. MiRNAs that consistently ranked in the top 10 across all 10 methods were chosen as the primary Biomarker Panel. We also identified a secondary panel comprising the top 10 miRNAs selected by at least 8 methods.

2.3.2. Evaluate Performance and External Validation

Predictive models were then built based on NB and RF algorithms using the selected miRNAs to evaluate the performance of the two biomarker panels on the GSE106817 and independent GSE164174 datasets.

Performance was evaluated using 10-fold cross-validation with five metrics: Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). The first four metrics were calculated from the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) through specific formulas. Additionally, the AUC was derived from a Receiver Operating Characteristic (ROC) curve.

External validation of the GSE164174 dataset was performed using cross-validation. The following parameters were set for the two classifiers of this study:

RF: Number of trees: 100; Criterion: gain_ratio; Maximal depth: 10; Voting strategy: confidence vote; Guess subset ratio: accurate.

NB: Laplace correlation parameter was set to true.

2.4. Cellular Verification of the Circulating 3-miRNA Signature in GC using scRNA-seq

We analyzed five tumor samples from the GC dataset STAD_GSE167297 through the Tumor Immune Single-Cell Hub 2 (TISCH2) database. Processed data and cell-type annotations were obtained from TISCH2, with the annotation resolution set to "major lineage." The localization and expression patterns of target genes corresponding to our 3-miRNA signature across different TME cell types were visualized using violin plots and feature plots.

To functionally characterize the primary cellular sources of the 3-miRNA signature, we performed Gene Set Enrichment Analysis (GSEA) on the same scRNA-seq dataset (STAD_GSE167297) used for cellular localization. Using the TISCH2 platform, GSEA was performed by comparing the five tumor samples with their paired normal tissues, focusing on the primary cell types that express the miRNA target genes. This analysis identified KEGG pathways that were significantly enriched among upregulated and downregulated gene sets within these cell populations. These cell-specific pathway enrichment results were subsequently compared to bulk analysis findings to determine cellular contributors to the bulk transcriptomic signatures.

2.5. Drug repurposing

To identify potential therapeutic agents for gastric cancer, a drug-gene interaction analysis was performed using DGIdb. This open-access resource aggregates records of genes and gene products, drugs, and drug-gene interactions to aid hypothesis generation and discovery for both clinicians and researchers [16]. All FDA-approved drugs known to influence genes targeted by the 5-miRNA Blood Signature were included. Additionally, a miRNA–target gene–drug interaction network was constructed using Cytoscape (version 3.10.1) [17] To visualize potential drugs and key target genes.

3. Results

3.1. Step 1: Specific DEMs Discovery

Principal component analysis (PCA) was performed on all five datasets for quality assessment prior to differential expression analysis (Figure S1 (Online Resource 1)). In the tumor tissue datasets, GSE54129 exhibited 7,013 DEGs, while GSE113255 showed 7,231 DEGs (Tables S1–S2 (Online Resource 2); Figures S2(Online Resource 1)). Figure S3 (Online Resource 1) shows the heatmaps of the most significant DEGs across all datasets in this study. A total of 2,473 common DEGs were identified for pathway enrichment analysis of the tissue datasets, resulting in 103 specific tissue pathways (Tables S3–S4 (Online Resource 2). The top 20 most significant pathways are illustrated in Fig. 2A. For the circulating miRNA

dataset (GSE106817), 2,010 DEMs were identified, with 2,925 target genes extracted from miRTarBase (Tables S5–S6 (Online Resource 2); Figure S2C). By intersecting these targets with the 2,519 plasma-derived DEGs from GSE174302 (Figure S2D), we identified 384 validated miRNA target genes in the blood of GC patients (Tables S7-S8 (Online Resource 2), Fig. 2B). Pathway enrichment analysis of these target genes identified 200 tissue-specific pathways (Table S9 (Online Resource 2)). The 20 most significant pathways appear in Fig. 2C.

Intersecting these 200 pathways with those 103 from tissue-derived DEGs resulted in 39 common pathways (Fig. 2D), ultimately mapping 59 miRNAs as specific DEMs (Table S10 (Online Resource 2)). These 39 pathways, along with their corresponding DEMs, are presented in Fig. 2E. These circulating DEMs were found to have strong mechanistic links to GC.

3.2. Step 2: Biomarker Panel Selection and Validation

Our selection strategy, which identified miRNAs ranked in the top 10 by at least 8 feature selection methods, yielded a biomarker panel comprising hsa-miR-124-3p, hsa-miR-23a-3p, hsa-miR-22-3p, hsa-miR-29b-3p, and hsa-miR-92a-3p. We evaluated this panel by constructing NB and RF models using the GSE106817 dataset (the same dataset employed in step 1 and feature selection). The RF classifier demonstrated superior performance (SEN: 77.39%; SPE: 99.24%; ACC: 98.36%; AUC: 98.50%; MCC: 78.28%), slightly outperforming NB.

To further validate the panel, we applied the RF algorithm to the five selected miRNAs on an independent dataset (GSE164174) that was not used in step 1 or the initial feature selection. This model also achieved high performance metrics: SEN (93.79%), SPE (84.69%), ACC (89.24%), AUC (95.30%), and MCC (78.80%) (Table 2; Fig. 3A, B).

Table 2
Performance metrics (expressed as percentages) for the RF-based miRNA biomarker panels in blood miRNA datasets.

MiRNA panels	Dataset ID	SEN	SPE	ACC	MCC	AUC
5-miRNA Panel	GSE106817	77.39	99.24	98.36	78.28	98.50
	GSE164174	93.79	84.69	89.24	78.80	95.30
	GSE106817	76.52	98.99	98.09	75.19	98.40
3-miRNA Panel	GSE164174	91.88	83.06	87.47	75.24	94.40

Notably, three miRNAs, hsa-miR-124-3p, hsa-miR-23a-3p, and hsa-miR-22-3p, were ranked among the top 10 by all 10 feature selection methods. Predictive models built using these miRNAs were evaluated, with RF again showing the highest performance (SEN: 76.52%; SPE: 98.99%; ACC: 98.09%; MCC: 75.19%; AUC: 98.40%), marginally surpassing NB.

External validation of the 3-miRNA panel via RF on the GSE164174 dataset further confirmed robust performance (SEN: 91.88%; SPE: 83.06%; ACC: 87.47%; AUC: 94.40%; MCC: 75.24%) (Table 2; Fig. 3C, D).

Performance evaluation through 10-fold cross-validation confirmed the reliability of the biomarker panel on the external validation dataset, supporting its potential as a diagnostic biomarker panel for gastric cancer.

3.3. Cellular Verification of the 3-miRNA Signature in GC using scRNA-seq

Unsupervised clustering of 22,464 cells revealed 17 distinct transcriptional clusters visualized via UMAP (Fig. 4A). To assign biological identities to these clusters, TISCH2 analyzed canonical marker gene expression patterns, which enabled annotation of nine major cell types (Fig. 4B). These included B cells, CD8 T cells, dendritic cells, endothelial cells, epithelial cells, fibroblasts, mast cells, monocytes/macrophages, and plasma cells. The relative abundance of each cell type across individual GC patients is shown in a bar chart (Fig. 4C). The global cellular landscape, representing the proportion of all cell types in the cohort, is summarized in a pie chart (Fig. 4D).

The TISCH2 database was utilized to visualize the expression of the 3-miRNA signature target genes through violin plots and feature plots across multiple cell types. Violin and feature plots revealed cell-type-specific expression patterns for the miRNA target genes (Fig. 4E-H). The mean expression of targets for hsa-miR-124-3p, RDH10, and CEBPA was highest in fibroblasts. Similarly, the mean expression of hsa-miR-22-3p targets (ESR1, PPARA, and BMP7) was primarily localized to fibroblasts and mast cells. The target of hsa-miR-23a-3p, HES1, showed broad expression across endothelial cells, fibroblasts, epithelial cells, and mast cells, in descending order. An integrated violin plot of the mean expression for all targets of the three miRNAs identified endothelial cells, fibroblasts, and epithelial cells as the principal cellular sources of the combined signature.

Accordingly, GSEA was performed on gene expression data from endothelial, epithelial, and fibroblast cell subsets. The enrichment patterns closely recapitulated the core biological processes identified in the bulk analysis. In the tumor microenvironment, upregulated DEGs across all three cell types (endothelial, epithelial, and fibroblast) were enriched in KEGG pathways related to key processes (including immune system regulation, hemostasis, signal transduction, and ECM organization), which corresponded to the Reactome pathways identified from the bulk GC data (39 shared pathways between GC tissue and blood samples). Simultaneously, downregulated DEGs in epithelial cells and fibroblasts were significantly enriched in several immune system pathways. Together, these results validate the pathway findings from the bulk analysis (Figs. 4I, 4J, and 6).

3.4. Drug-Gene Interaction Identification for the 5-miRNA Signature Targets

To identify candidate drugs for repurposing, we investigated drug–gene interactions for the target genes of the 5-miRNA signature. Querying the DGIdb server identified 143 FDA-approved drugs interacting with 8 of these target genes, confirming their high drugability. Specifically, analysis revealed 106 candidate drugs for ESR1, 20 for PPARA, and 10 each for BMP7 and DNMT3A. Four target genes (DNMT3B, ITGA5, HES1, and S100B) were each linked to only a single drug candidate. Notably, four drugs (including Olaparib, Doxorubicin Hydrochloride, Bithionol, and Daunorubicin Liposomal) interacted with more than one target gene, highlighting their potential for further investigation in GC treatment (Fig. 5).

4. Discussion

This study presents a novel, integrated bioinformatics and machine learning pipeline that successfully identifies a highly specific, dual-purpose 5-miRNA signature in blood for gastric cancer. Our approach, which roots biomarker discovery in the fundamental biology of the tumor, moves beyond correlative associations to establish a mechanistic link between circulating signals and tissue pathology. The signature demonstrates exceptional diagnostic performance and, through subsequent analysis, reveals a network of potential therapeutic targets, offering a comprehensive strategy to address the critical challenges of late diagnosis and limited treatment options in GC. The primary innovation of our methodology lies in the systematic integration of transcriptomics data to enforce biological specificity. Although numerous studies have catalogued differentially expressed circulating miRNAs in GC, an unclear mechanistic link to gastric tumorigenesis have limited their translational impact [18, 19]. We hypothesized that circulating miRNAs reflective of GC biology could be identified by first finding pathways common to both GC tumor tissue and blood, and then finding the associated miRNAs present in the blood. By first identifying shared dysregulated pathways between tissue-derived mRNA (GSE54129, GSE113255) and the target genes of blood-derived miRNAs (GSE106817, refined by intersection with blood mRNA data GSE174302), we created a powerful filter. This pathway-centric filtering strategy yielded 59 miRNAs whose regulatory activity is directly pertinent to GC-specific processes, including ECM organization, signal transduction, immune system regulation, and hemostasis (Fig. 6). This design distinguishes our work from purely statistical or machine-learning-driven biomarker discovery, ensuring that the final panel is not merely a statistical artifact but a reflection of core pathogenic mechanisms. For instance, the enrichment of ECM organization pathways aligns with established knowledge of its role in GC development and progression, providing a plausible biological context for the selected miRNAs [20]. From the biologically curated set of 59 miRNAs, we employed a robust ensemble machine-learning strategy to derive a minimal yet maximally informative diagnostic panel. By integrating eight filter methods (Gini index, information gain, information gain ratio, chi-squared, uncertainty, deviation, correlation, Relief) and two embedded approaches (tree importance, rule-based weights), we mitigated algorithm-specific bias and identified a consensus 5-miRNA panel: hsa-miR-124-3p, hsa-miR-23a-3p, hsa-miR-22-3p, hsa-miR-29b-3p, and hsa-miR-92a-3p. This panel achieved excellent diagnostic performance in the discovery cohort (GSE106817) with an RF classifier, achieving an AUC of 98.50% and an accuracy of 98.36%, and retained high performance in an independent large validation cohort (GSE164174; AUC 95.30%, accuracy 89.24%). These metrics compare favorably with,

and in many cases surpass, those reported in previous studies utilizing ML for GC miRNA biomarker discovery [21–23]. Furthermore, the panel's performance remained high even when reduced to a 3-miRNA core (hsa-miR-124-3p, -23a-3p, -22-3p), which was ranked top by all feature selection methods, enhancing its potential for cost-effective clinical translation. A critical question for circulating miRNA biomarkers concerns their cellular origin. To address this, we performed scRNA-seq analysis on gastric tumor tissue, enabling high-resolution mapping of target gene expression within the TME. The analysis revealed a predominant stromal signature. Targets of hsa-miR-124-3p (RDH10, CEBPA) and hsa-miR-22-3p (ESR1, PPARA, BMP7) were most highly expressed in fibroblasts, with miR-22-3p targets also enriched in mast cells. In contrast, the miR-23a-3p target HES1 was broadly expressed across endothelial, epithelial, and immune cells. These results implicate a network of cells, with fibroblasts emerging as a central hub. This suggests that the circulating levels of these miRNAs may reflect a complex interplay between stromal, immune, and epithelial compartments. These findings provide a cellular context for the observed biomarker levels and offer new hypotheses regarding the functional roles of these miRNAs in the tissue microenvironment. Moreover, this scRNA-seq analysis identified fibroblasts, endothelial cells, and epithelial cells as the most promising cellular targets for future therapeutic interventions. GSEA at the single-cell level recapitulated the same pathway themes (e.g., immune regulation, ECM organization, hemostasis, signal transduction) identified in the bulk analysis, creating a robust, multi-scale validation of the signature's biological coherence. Beyond diagnosis, our integrated pipeline facilitates the direct translation of biomarker discovery into therapeutic hypothesis generation. Accordingly, we extended our integrated pipeline to drug repurposing analysis based on the target genes of the 5-miRNA signature. ESR1 emerged as the most centrally located node, connecting to more than 100 drugs and the tumor-suppressive miRNA hsa-miR-22-3p, which has been reported to inhibit proliferation and invasion in gastric cancer [24]. This was followed by PPARA, BMP7, and DNMT3A, all of which are linked to multiple FDA-approved drugs (Fig. 5). Although ESR1 has been widely investigated in hormone-dependent cancers like breast and endometrial cancer, there is also evidence linking it to the development of gastric cancer progression, where altered ESR1 signaling has been linked to enhanced proliferation, invasion, and poor prognosis [25, 26]. Network analysis identified a potential miR-22-3p–ESR1 axis, consistent with miR-22's tumor-suppressive role in gastric cancer [24]. Interestingly, drugs including Olaparib, Doxorubicin hydrochloride, Daunorubicin liposomal, and Bithionol interacted with multiple key targets (ESR1, PPARA, BMP7, and DNMT3A), suggesting overlapping regulatory mechanisms. Notably, Olaparib showed the highest interaction confidence with ESR1, suggesting a potential regulatory link between estrogen receptor signaling and DNA repair–related drug pathways in GC. The prominent presence of DNA-damaging agents such as Doxorubicin, along with prior evidence of synergy between Olaparib and Doxorubicin [27–29], suggests that similar regulatory or synergistic interactions may operate in gastric cancer, potentially mediated by miRNA-driven modulation of DNA repair pathways. The co-occurrence of ESR1, Olaparib, Doxorubicin, and miR-22-3p within the same interaction module suggests potential miRNA–ESR1–drug axis involved in gastric cancer's response to chemotherapy. (Fig. 5). In parallel, both drugs are also associated with DNMT3A, a DNA methyltransferase targeted by miR-29b-3p. DNMT3A is implicated in gastric cancer development through epigenetic regulation, and suppression of DNMT3A by the miR-29 family has been associated

with reduced tumor proliferation and metastatic potential [30], suggesting a complementary miR-29b-3p/DNMT3A/drug axis that potentially links DNA methylation dynamics with DNA damage–repair responses in gastric cancer. PPARA and BMP7 were identified as additional drug-associated targets regulated by miR-22-3p. PPARA, previously reported to be upregulated in GC cell lines [31, 32], was linked to Bithionol and clustered with ESR1 in the interaction network, suggesting a possible crosstalk between PPARA-mediated metabolic regulation and ESR1-driven hormone receptor signaling. BMP7 is involved in the regulation of apoptosis, epithelial–mesenchymal transition, and DNA damage response pathways, which have been reported to modulate cellular responses to DNA-damaging agents [33]. BMP7 expression has been observed not only on the cell membrane but also in the cytoplasm of gastric cancer cells. It is significantly associated with nodal involvement, lymphatic invasion, and venous invasion [34]. Given the apoptosis-inducing mechanisms of anthracyclines such as Doxorubicin and Daunorubicin [35], the network-based association of BMP7 with miR-22-3p and anthracyclines suggests a potential regulatory axis (miR-22-3p/BMP7/anthracycline) that may modulate chemosensitivity in gastric cancer, although direct experimental evidence remains limited.

Despite the strengths of this integrated approach, certain limitations must be acknowledged. First, while we used multiple independent datasets for discovery and validation, all analyses are retrospective. Prospective validation in a multi-center clinical trial, ideally collecting matched tissue, blood, and clinical outcome data, is the essential next step to confirm the panel's real-world diagnostic and prognostic utility. Second, the mechanistic predictions from our network analysis, (particularly the proposed miR-22-3p/ESR1/Olaparib and miR-29b-3p/DNMT3A/Doxorubicin axes) require direct experimental validation using *in vitro* and *in vivo* GC models. Functional assays are needed to confirm the regulation of target genes by these miRNAs and to test the synergy or resistance mechanisms with the identified drugs. Third, while scRNA-seq identified likely cellular sources, the precise vesicular packaging and release mechanisms (e.g., exosomes, microvesicles) by which these miRNAs are secreted from TME cells into the circulation remain to be elucidated. Finally, the clinical feasibility of measuring a 5-miRNA panel in a standardized, cost-effective assay format needs to be developed and optimized.

Conclusion

We developed an integrated workflow combining pathway-based analysis, machine learning, and single-cell sequencing. This allowed the discovery of circulating blood biomarkers and therapeutic targets for gastric cancer. Our approach identified two biologically grounded miRNA signatures that demonstrated diagnostic performance and were validated by machine-learning and scRNA-seq analyses. Single-cell profiling confirmed the biological origins of the tumor signatures. Network-based drug repurposing analysis provided mechanistic insight into how these miRNAs may relate to therapeutic response. ESR1 emerged as a pivotal regulatory target, linking tumor-suppressive miR-22-3p to various chemotherapeutic agents, including Olaparib and Doxorubicin. The identified miR-22-3p–ESR1–PARPi axis suggests a mechanism by which estrogen receptor signaling may affect DNA repair and drug response in gastric cancer. Concurrently, the miR-29b-3p/DNMT3A and miR-22-3p/PPARA/BMP7 pathways suggest a wider miRNA-mediated regulation of epigenetic and apoptotic processes. Taken

collectively, these results uncover a unified miRNA-gene-drug regulome that combines diagnosis with the generation of a hypothesized therapy for therapy. Even if further validation in the clinic will be needed, a basis for a personalized approach to gastric cancer treatment might be established using such a regulome.

Electronic supplementary material: The online version of this article contains supplementary material, including Figures S1-S3 and Tables S1–S10, available to authorized users (Online Resource 1 and Online Resource 2), respectively.

Declarations

Acknowledgments: The authors thank the editor and reviewers for their valuable feedback, which greatly enhanced the manuscript.

Authors' contribution**:** A.G. and M.M. conceived and designed the study. S.H., M.S., S.M., S.J, M.M., and S.S. were responsible for data acquisition and analysis. S.H. and A.G. interpreted the results. Figure generation was performed by S.H., M.M., M.S., S.J., and S.M., while tables were generated by S.J., S.M., S.H., and M.M. The manuscript was drafted by M.M., S.H., A.G., M.S., and S.M. All authors reviewed and approved the final manuscript.

Funding source:This study has no funding sources.

Data availability statement: No human data are directly involved in the present study's analysis. The data described in this article can be freely and openly accessed at the GEO database on the following links:

GSE54129 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54129],

GSE113255 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113255],

GSE174302 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174302],

GSE106817 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106817],

GSE164174 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164174],

STAD_GSE167297 [https://tisch.compbio.cn/gallery/?cancer=STAD&celltype=&species=].

Conflict of Interest: The authors declare no conflicts of interest in association with the present study.

Ethics approval: Ethical approval was not required for this study as it did not involve any experiments on human participants or animals.

Consent to Publish declaration: not applicable

Consent to Participate declaration: not applicable

Ethics approval and consent to participate: Not applicable. This study is a secondary analysis of publicly available, fully de-identified human gastric cancer datasets obtained from open-access repositories, including The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). No new human tissues, clinical samples, or experimental interventions were collected or performed by the authors. Therefore, informed consent was not obtained directly from participants, and no institutional review board or ethics committee approval was required specifically for this bioinformatics analysis. However, we confirm that all original studies that generated and deposited these datasets were conducted in full accordance with the ethical principles of the Declaration of Helsinki and obtained appropriate informed consent and institutional ethical approval at the time of data collection. All analyses performed in the current study strictly adhered to the guidelines for the responsible and ethical use of publicly available human data

References

1. Sung H, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Ajani JA, et al. Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2022;20(2):167–92.
3. Carpelan-Holmström M, et al. CEA, CA 19 – 9 and CA 72 – 4 improve the diagnostic accuracy in gastrointestinal cancers. Anticancer Res. 2002;22(4):2311–6.
4. Song Z, et al. Progress in the treatment of advanced gastric cancer. Tumour Biol. 2017;39(7):1010428317714626.
5. Gatfield ER, Tadross J, Ince W. Immune checkpoint inhibitor use in head and neck squamous cell carcinoma: the current landscape and future perspectives. Future Oncol. 2024;20(23):1695–711.
6. Sharifı Z, Talkhabi M, Taleahmad S. Identification of potential microRNA diagnostic panels and uncovering regulatory mechanisms in breast cancer pathogenesis. Sci Rep. 2022;12(1):20135.
7. Roointan A, et al. Candidate MicroRNA Biomarkers in Lupus Nephritis: A Meta-analysis of Profiling Studies in Kidney, Blood and Urine Samples. Mol Diagn Ther. 2023;27(2):141–58.
8. Gholaminejad A, et al. A meta-analysis of microRNA expression profiling studies in heart failure. Heart Fail Rev. 2021;26(4):997–1021.
9. Gholaminejad A, Abdul Tehrani H, Gholami M, Fesharaki. Identification of candidate microRNA biomarkers in renal fibrosis: a meta-analysis of profiling studies. Biomarkers. 2018;23(8):713–24.
10. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic - implementation issues and future challenges. Nat Rev Clin Oncol. 2021;18(5):297–312.
11. Azari H, et al. Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer. Sci Rep. 2023;13(1):6147.

12. Masoudi-Sobhanzadeh Y, et al. Discovering driver nodes in chronic kidney disease-related networks using Trader as a newly developed algorithm. Comput Biol Med. 2022;148:105892.
13. Momeni M, et al. A comprehensive analysis of gene expression profiling data in COVID-19 patients for discovery of specific and differential blood biomarker signatures. Sci Rep. 2023;13(1):5599.
14. Zhu SL, et al. Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics. PLoS ONE. 2020;15(12):e0244869.
15. Gilani N, et al. Identifying Potential miRNA Biomarkers for Gastric Cancer Diagnosis Using Machine Learning Variable Selection Approach. Front Genet. 2021;12:779455.
16. Cannon M, et al. DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. Nucleic Acids Res. 2024;52(D1):D1227–35.
17. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
18. Huang Z, et al. Six serum-based miRNAs as potential diagnostic biomarkers for gastric cancer. Cancer Epidemiol Biomarkers Prev. 2017;26(2):188–96.
19. Liu R, et al. A five-microRNA signature identified from genome-wide serum microRNA expression profiling serves as a fingerprint for gastric cancer diagnosis. Eur J Cancer. 2011;47(5):784–91.
20. Moreira AM, et al. The extracellular matrix: an accomplice in gastric cancer development and progression. Cells. 2020;9(2):394.
21. So JBY, et al. Development and validation of a serum microRNA biomarker panel for detecting gastric cancer in a high-risk population. Gut. 2021;70(5):829–37.
22. Sui S, et al. Exosomal liquid biopsy for the early detection of gastric cancer: the DESTINEX multicenter study. JAMA Surg. 2025;160(9):973–82.
23. Ma F-C, et al. Blood-based machine learning classifiers for early diagnosis of gastric cancer via multiple miRNAs. World J Gastrointest Oncol. 2025;17(4):103679.
24. Wang X, et al. MiR-22 suppresses the proliferation and invasion of gastric cancer cells by inhibiting CD151. Biochem Biophys Res Commun. 2014;445(1):175–9.
25. Shang L, et al. Cancer-Associated Fibroblast-Secreted Exosomes Promote Gastric Cancer Cell Migration and Invasion via the IL-32/ESR1 Axis. Appl Biochem Biotechnol. 2024;196(9):6045–58.
26. Tang W, et al. Expression of estrogen receptors and androgen receptor and their clinical significance in gastric cancer. Oncotarget. 2017;8:40765–77.
27. Wang Y, et al. PARP inhibitors in gastric cancer: beacon of hope. J Exp Clin Cancer Res. 2021;40(1):211.
28. Kciuk M et al. Doxorubicin-An Agent with Multiple Mechanisms of Anticancer Activity. Cells, 2023. 12(4).
29. Park HJ, et al. The PARP inhibitor olaparib potentiates the effect of the DNA damaging agent doxorubicin in osteosarcoma. J Exp Clin Cancer Res. 2018;37(1):107.

30. Cui H, et al. Deregulation between miR-29b/c and DNMT3A is associated with epigenetic silencing of the CDH1 gene, affecting cell migration and invasion in gastric cancer. PLoS ONE. 2015;10(4):e0123926.
31. Wang CY, et al. Upregulation of peroxisome proliferator-activated receptor-α and the lipid metabolism pathway promotes carcinogenesis of ampullary cancer. Int J Med Sci. 2021;18(1):256–69.
32. Jia Q, et al. Comprehensive analysis of peroxisome proliferator-activated receptors to predict the drug resistance, immune microenvironment, and prognosis in stomach adenocarcinomas. PeerJ. 2024;12:e17082.
33. Koganti P, Levy-Cohen G, Blank M. Smurfs in Protein Homeostasis, Signaling, and Cancer. Frontiers in Oncology; 2018. pp. 8–2018.
34. Aoki M, et al. Expression of BMP-7 in human gastric cancer and its clinical significance. Br J Cancer. 2011;104(4):714–8.
35. Shandilya MD, et al. Molecular-Level Understanding of the Anticancer Action Mechanism of Anthracyclines. Advances in Precision Medicine Oncology; 2020.

Figures

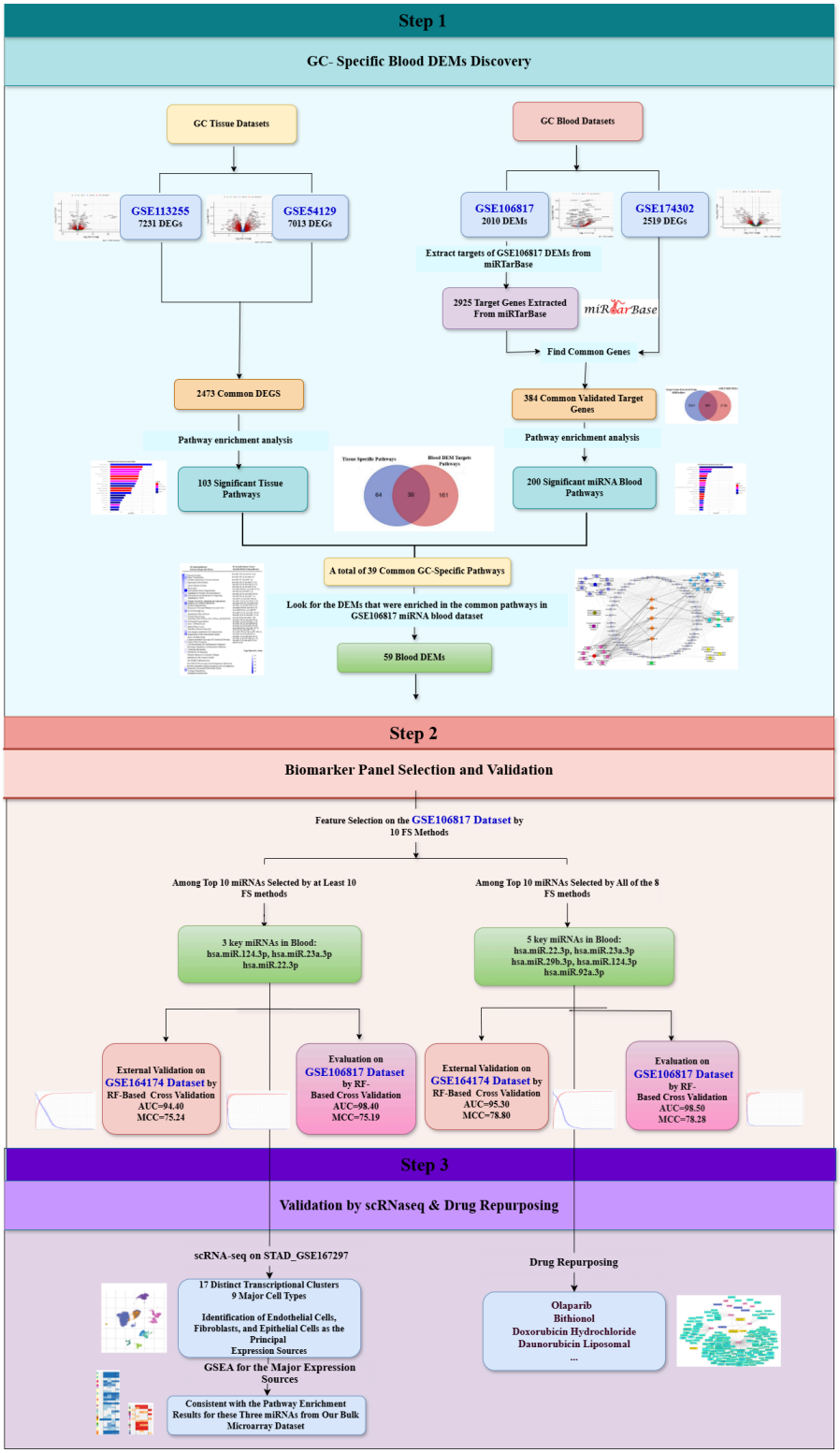


Figure 1

A workflow illustrating the key stages of the current research.

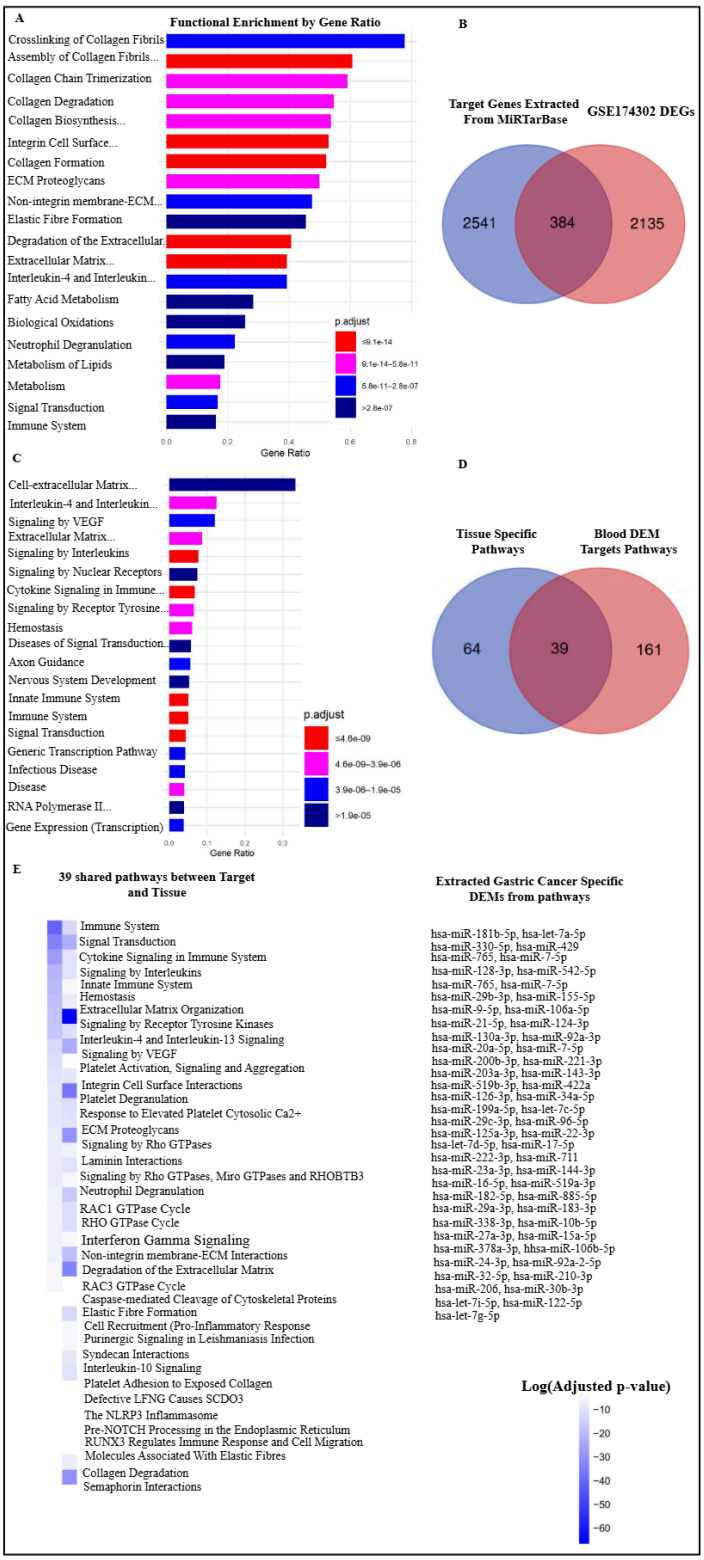


Figure 2

Bar plots show the top 20 most significant pathways for **A**, tissue datasets, and **C**, blood miRNA dataset. Venn diagram representing the overlapping genes between miRTarbase target genes and GSE174302 DEGs; **B**. Venn diagram representing the number of GC tissue and circulating miRNA target genes pathways; **D, E.** Common pathways shared between GC tissue and the circulating miRNA target genes,

along with their adjusted p-values from the pathway enrichment analysis, and the compilation of extracted Specific DEMs derived from these pathways.

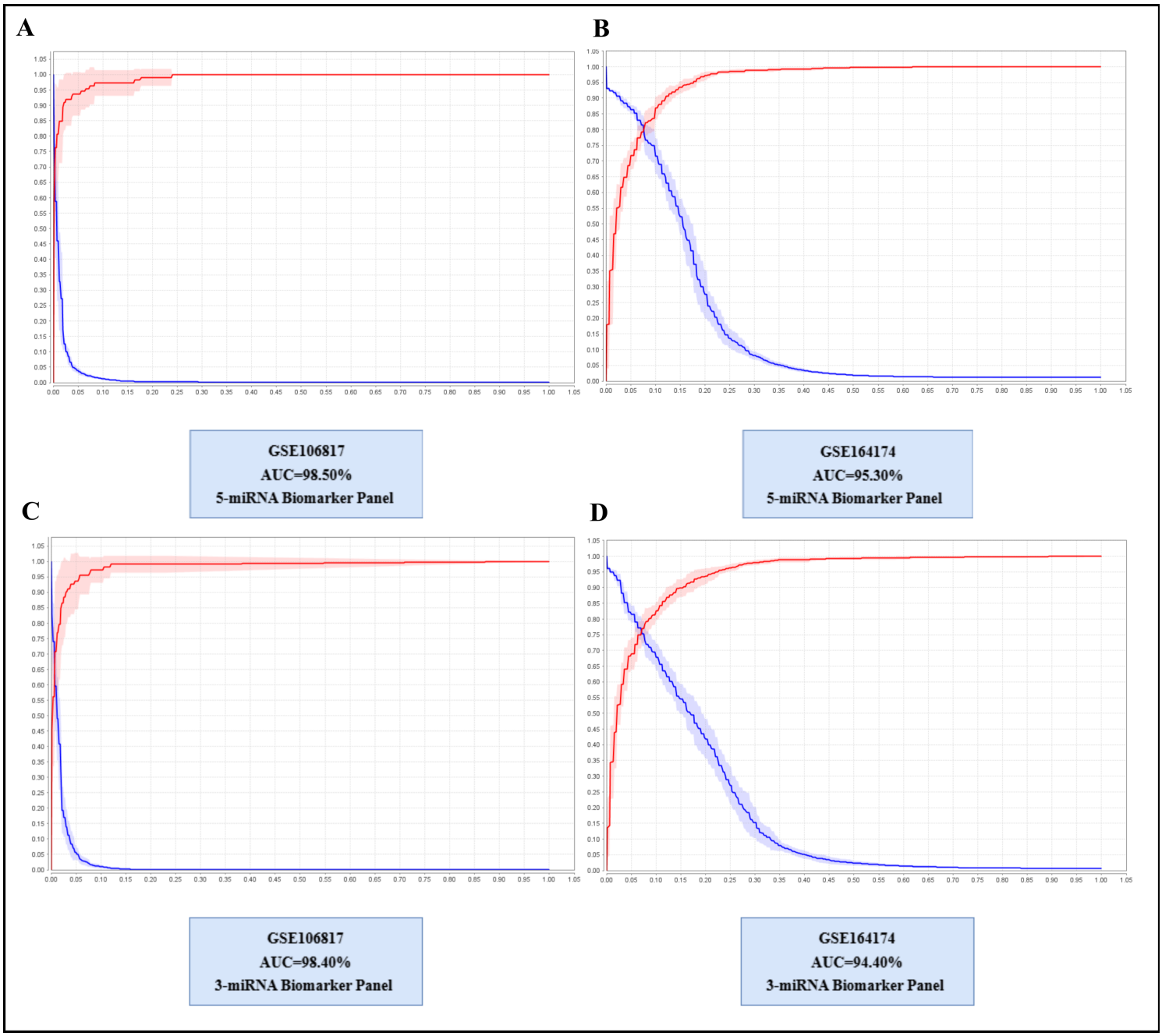


Figure 3

ROC analysis of the Random Forest classifier using the 5-miRNA and 3-miRNA biomarker panels. Shown are ROC curves for the 5-miRNA panel on the discovery dataset (GSE106817 (A and C)) and the external validation dataset (GSE164174 (B and D)), as well as corresponding ROC curves for the 3-miRNA panel on both datasets. Red lines represent the ROC curves, and blue lines show threshold variations. Axes indicate 1-specificity (x-axis) and sensitivity (y-axis).

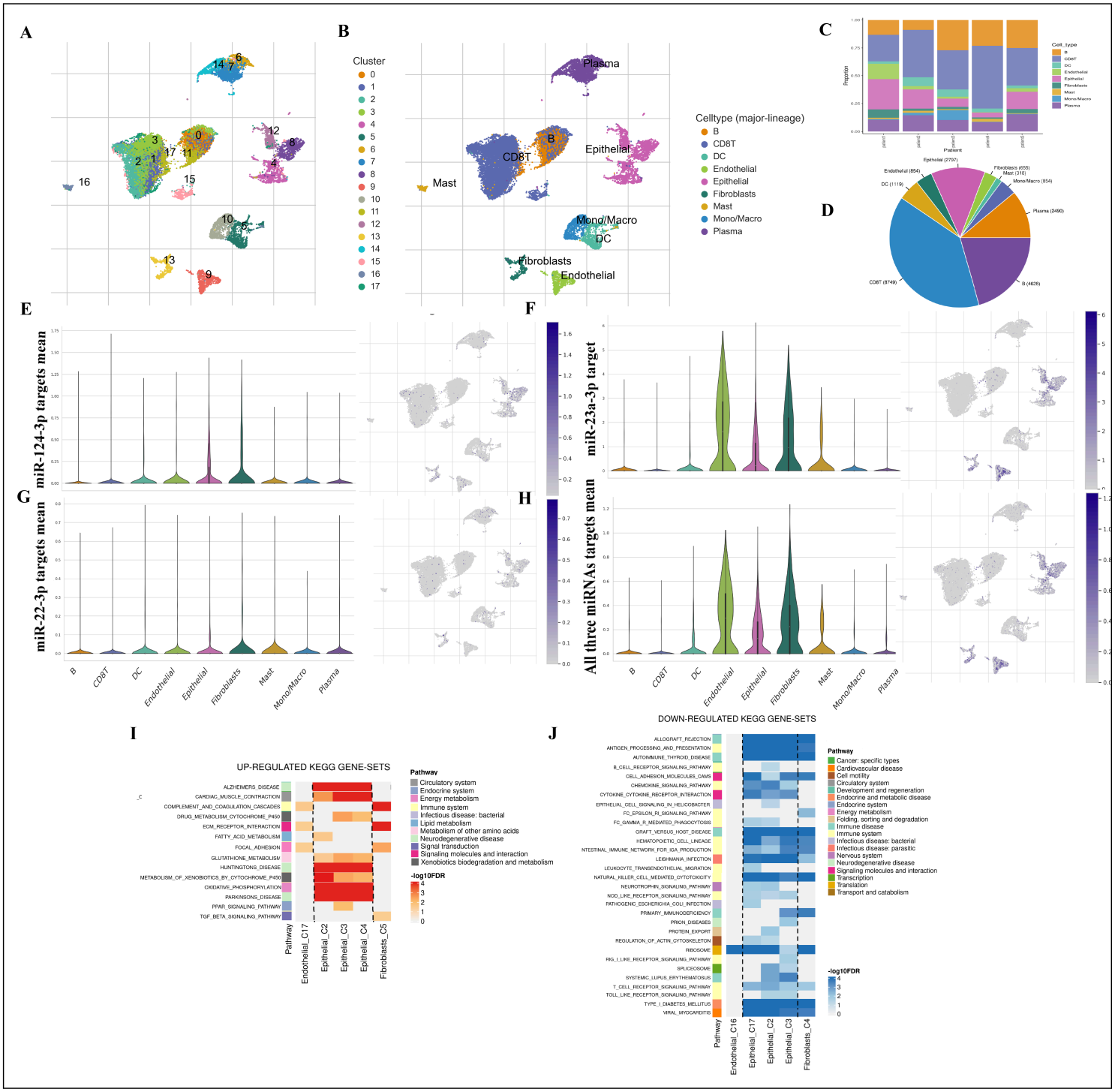


Figure 4

Single-cell transcriptomic profiling of gastric cancer reveals cellular heterogeneity, miRNA target localization, and pathway enrichment**. (A)** UMAP visualization of 17 transcriptionally distinct cell clusters. **(B)** Annotation of clusters into 9 major cell lineages. **(C)** Stacked bar chart showing cell type proportions across individual patients. **(D)** Pie chart summarizing overall dataset composition. **(E-G)** Violin and feature plots depicting mean expression of target genes for hsa-miR-124-3p **(E)**, hsa-miR-23a-3p **(F)**, and hsa-miR-22-3p **(G)**. **(H)** Combined expression profile for all targets from E-G. **(I, J)** GSEA

results showing significantly enriched KEGG pathways for upregulated **(I)** and downregulated **(J)** genes. Cell populations in GSEA are labeled as [cell type] C [cluster number].

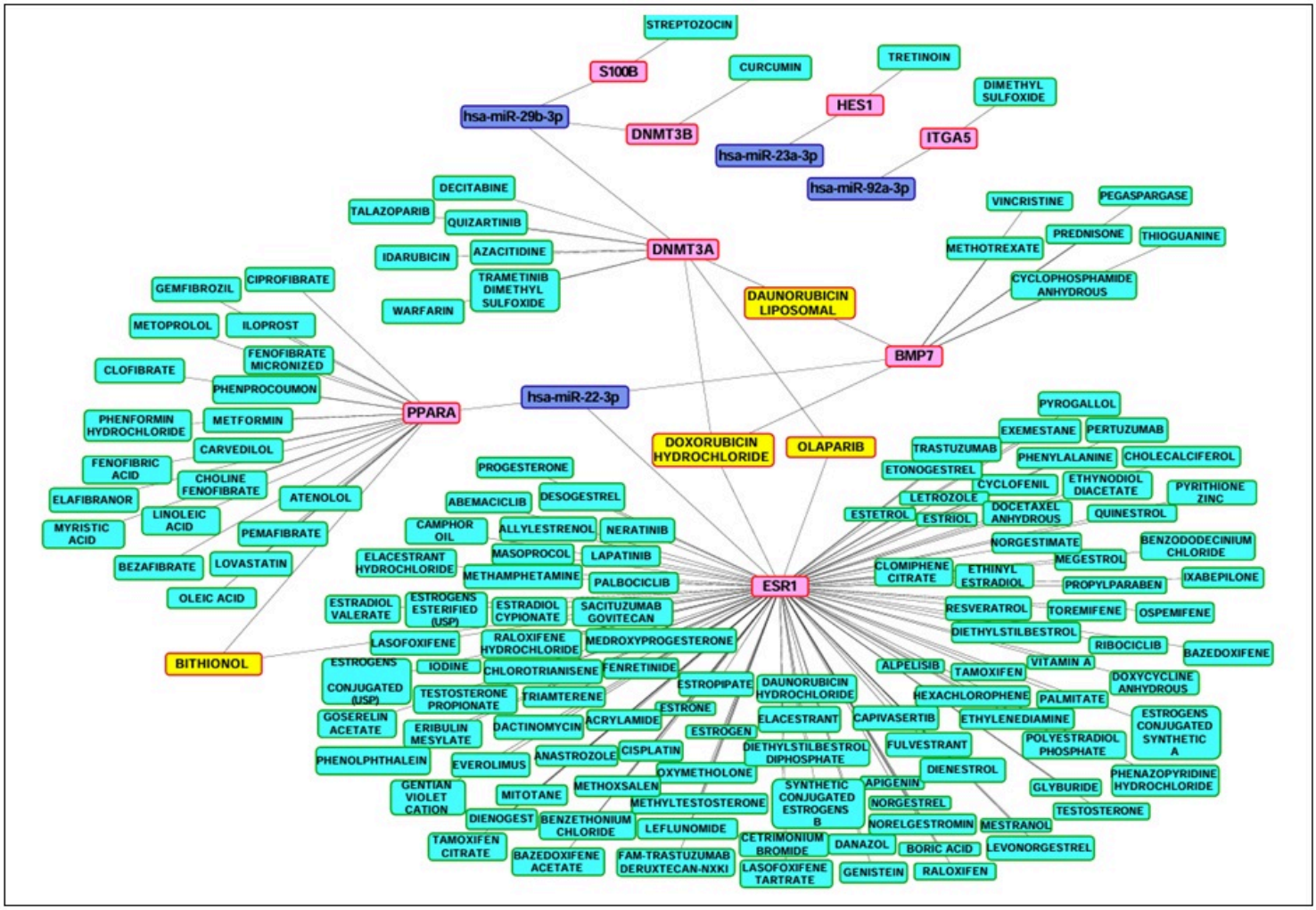


Figure 5

Network of potential drugs and target genes constructed in Cytoscape. In the network, bluish-purple nodes represent the 5 ML-validated miRNAs, pink nodes represent their target genes, and turquoise nodes represent potential drugs. Yellow highlights drugs that interact with multiple target genes.

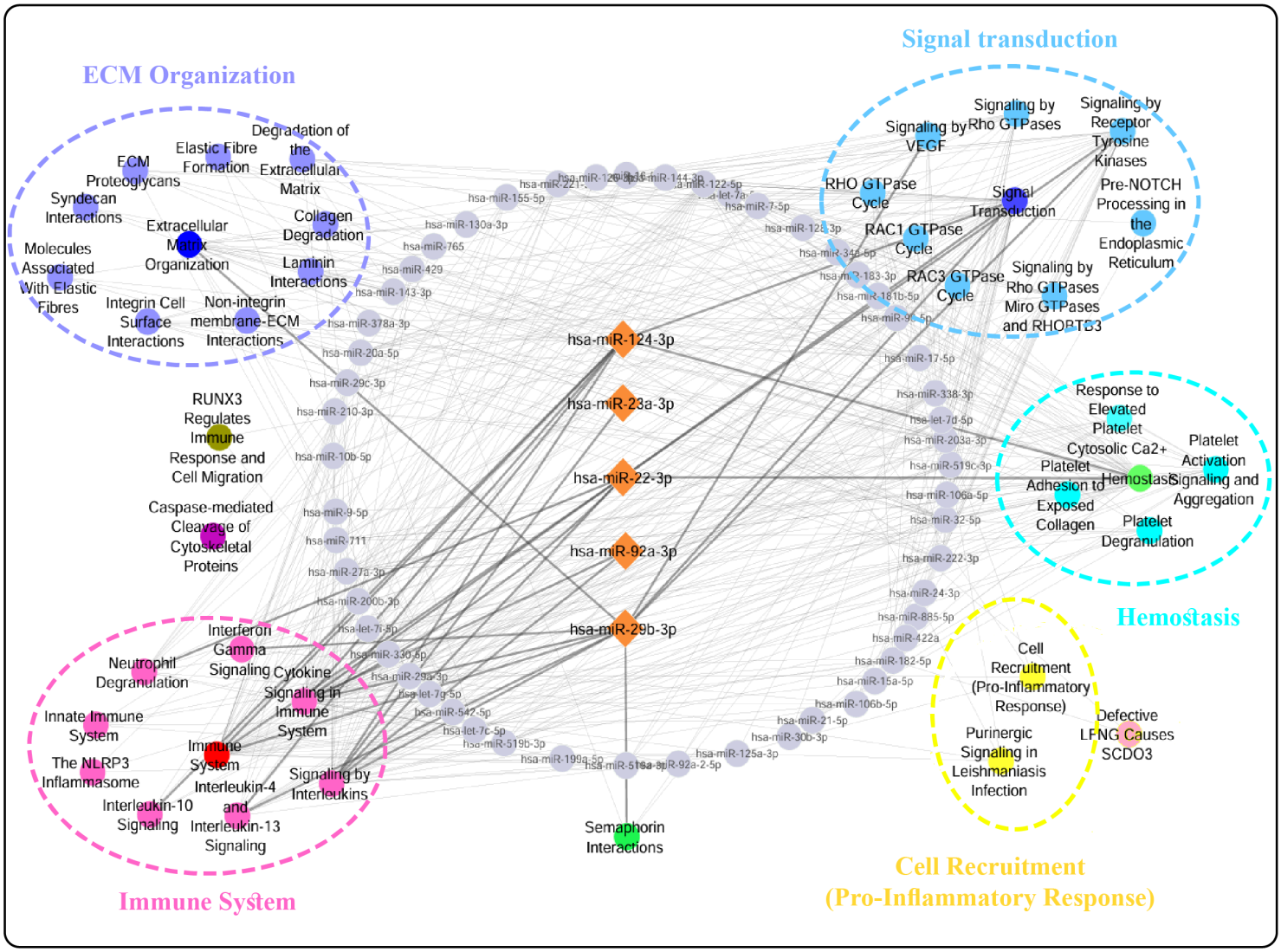


Figure 6

Interaction network of Reactom pathways shared between the GC tissue and blood miRNA datasets, showing DEMs enriched in these pathways in the blood dataset. The final 5-miRNA biomarker panel is shown at the center, with connections to its associated pathways.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1.pdf
- OnlineResource2.xlsx